Proposal for Tape Handling for CDF in Run 2

Robert M. Harris

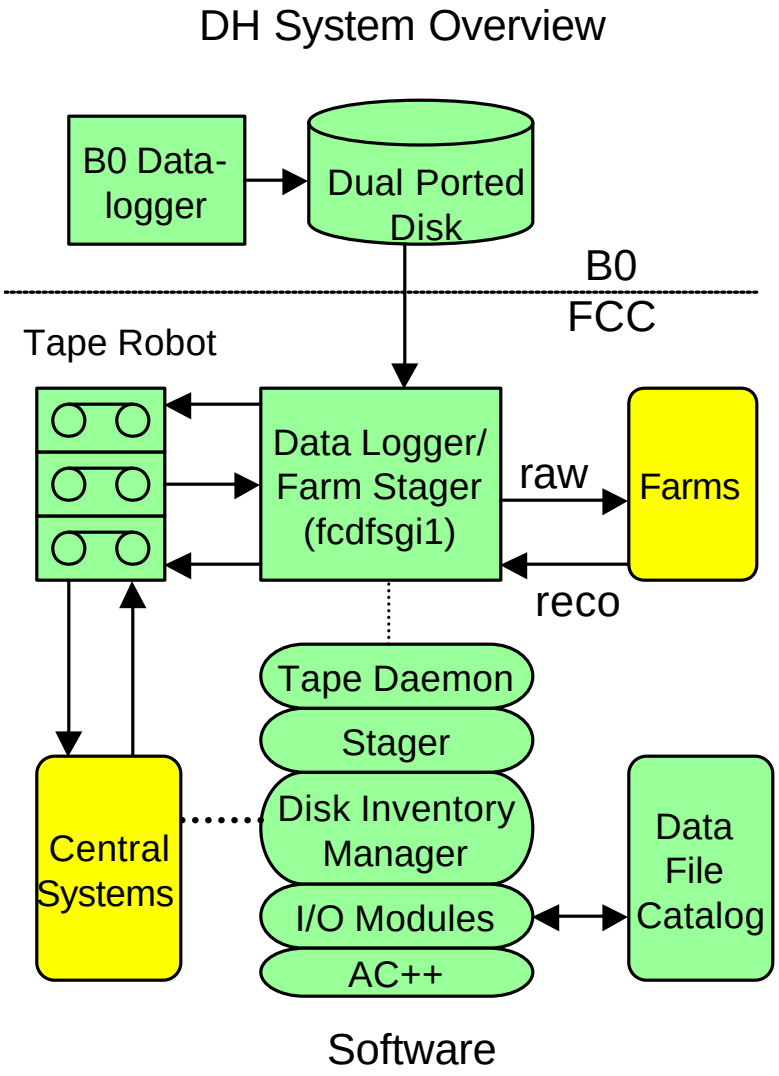
DH Review

Nov. 19, 2001

Outline

- Introduction to Proposal
  - Existing DH System
  - Problems with existing Tape Handling
  - Conceptual Description of Proposal
  - Benefits: support, operations, flexibility.
  - Mapping the Software Concepts
  - Serial Media Capacities & Costs
  - STK Silo Logistics
  - Staged Proposal
- Prototype Design & Schedule
  - Quick Prototype for Tape Handling
  - Run 1 Silo Migration Plans
- Proposal Variants
  - Network Issues
  - Network Design of Variants
  - Discussion of Variants
- Full Schedule
  - Installation Schedule
  - Use & Migration Options
- Conclusions

Existing Data Handling System



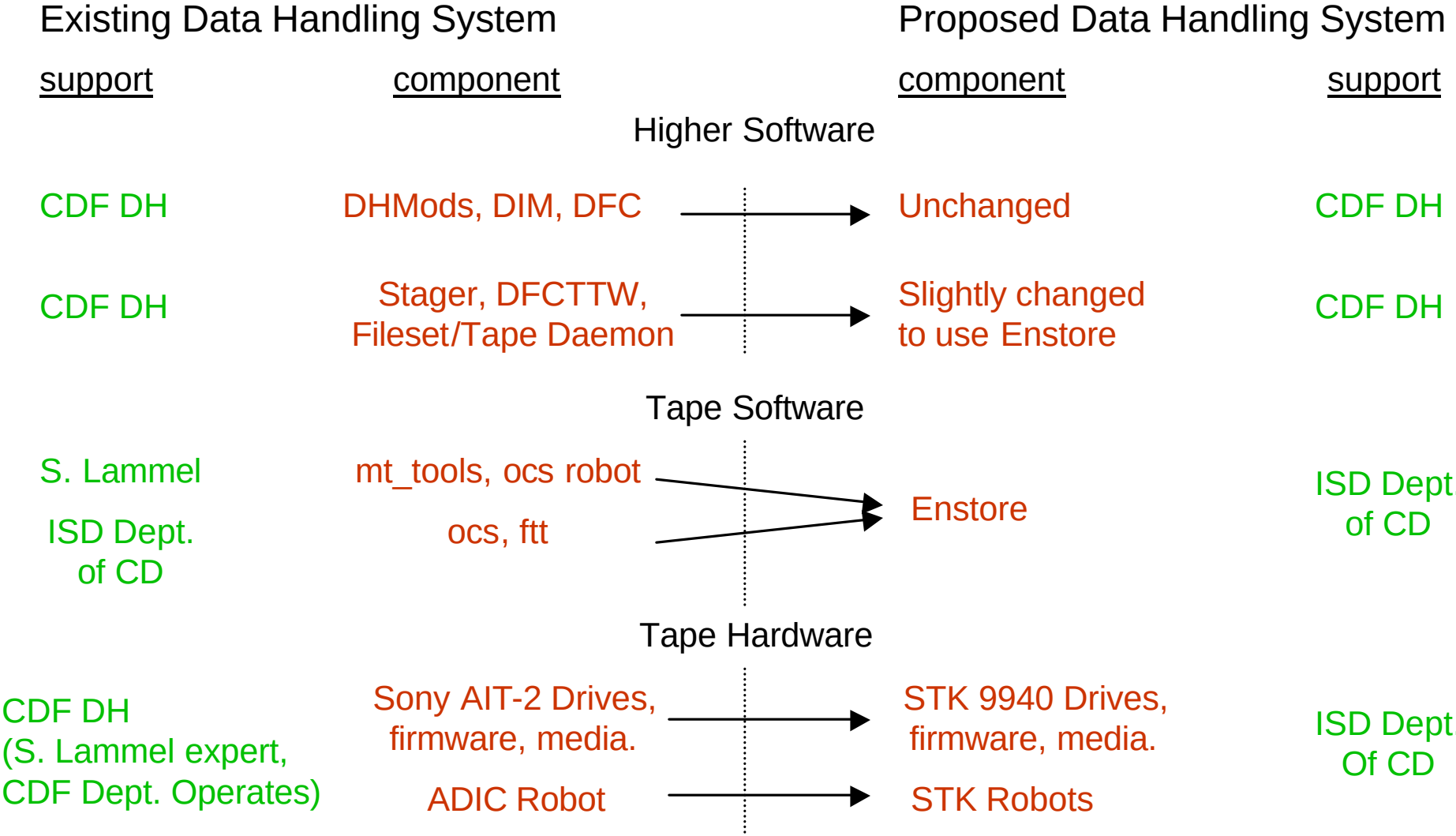
- DH Software Overview
  - DHMods: AC++ Input/Output Module
    - User Interface to analyzing data.
    - Talks to the DFC Oracle DB.
  - Data File Catalog
    - Saves metadata of files, filesets, tapes, datasets.
  - Disk Inventory Manager
    - Manages data on disk in filesets.
  - Stager
    - Reads filesets from tape to disk for DIM & farms.
  - Fileset / Tape Daemon
    - Writes raw / reco data to tape on fcdfsgi1.
    - Writes full tapes once all filesets are collected.
  - DFCTestTapeWrite
    - Writes secondary datasets to tape on fcdfsgi2.
  - mt_tools
    - Writes ANSI tapes & handles errors.
    - Supports tape partitioning & AIT-2 tapes.
  - OCS robot
    - Handles mounting of tapes in robot.
  - OCS & FTT
    - Tape allocating, reading, and writing primitives.

Problems with Existing Tape Handling

- Operational Problems
  - Problems use ~2 FTEs and hamper farms efficiency even now, without raw data taking.
    - Tape reading problems
      - **Short reads of tapes giving incomplete data on disk for farms & user analysis.**
      - **Tape reading problems of unknown origins which halt farms & user analysis.**
    - Tape dismounting problems
      - **Drives get deallocated but tape remains mounted in drive.**
      - **Job completes successfully but drive remains allocated and tape remains mounted.**
      - **Stuck tape queues caused by killed LSF jobs not dismounting tape drives.**
    - Tape writing problems
      - **More than 1% of all tape writes fail requiring intervention.**
    - Current system needs human attention every few hours!
  - The existing system will <u>not</u> be operable under more intense loads.

- Expertise & Support
  - Expert resigned leaving products & choices that are unsupported elsewhere.
    - Little expertise other than S. Lammel for CDF tape software & Sony AIT-2 tape drive firmware and hardware.

Conceptual Description of Proposal



Support & Operations Benefits

- CD Supported Tape Handling System at Fermilab
  - Used in the public STKEN robot by 16 experiments/groups.
    - E815, E872, Auger, Boone, BTeV, CKM, CMS, D0, E791, E831, KTeV, miniboone, Minos, SDSS, Selex, Theory.
  - Tape Handling system would be supported by ISD Department
    - Support drives in robot and robot operations like they do now for lab.
      - **Avoids CDF task force burning themselves out on unsupportable AIT-2 system.**
    - Support Enstore software for data handling like they do now for lab.
      - **Avoids wasted effort on unsupportable software used only by CDF.**
- Debugged tape handling software.
  - Solves our serious problems involving dismounts (stuck queues, etc.).
    - Enstore process that controls tapes is separate from user jobs.
    - If communication lost process times out and dismounts tapes.
- Reliable data center quality STK 9940 drives.
  - In service at FNAL, BNL, SLAC, CERN, DESY (9840 variant).
  - Write failures on less than 0.1% of tapes, compared to more than 1% for AIT-2.

Future Potential Benefits

- Central Analysis Facility
  - Pile of PC's proposed by CAF review gets data from same switch as SGIs.
    - Uses Stager & DIM that can already run on Linux.
      - **Unknown work to get direct attached tapes working on Linux in existing system.**
      - **Large quantity of additional direct attached drives to support in existing system.**
- Distribution to Trailers & Universities
  - Trailers could access data directly from the robot if desired.
    - Either through DH system on their desktop or via Enstore directly.
    - Priorities can be granted to raw data and farms in Enstore.
    - Link between Trailers & FCC might have to be upgraded and use policed.
  - We could still have a tape copy facility if desired.
    - Copy datasets onto AIT-2 tapes at FNAL and ship them to universities as planned.
- User archive
  - User and physics group archive of files to tape from their static data disks.
  - Storage of miscellaneous data files, root tuples, etc.
  - May reduce needed size of user and physics group static disk pool.
- Enstore and Network Access adds functionality and flexibility.

Mapping the Software Concepts

- Plan to map CDF DH to Enstore without changing either.
  - Enstore writes files to tapes
    - User does not control which tape the files go on.
  - Enstore doesn't support filesets which DH software uses.
    - ISD willing to consider adding these for the future, but meanwhile . . .
  - We plan to create "logical filesets" and "logical tapes".
    - Collections of files that are not necessarily on same physical tape.
    - User requests datasets, logical tapes, logical filesets, runs or files.
      - **Same as before only files may be spread out over multiple tapes.**
      - **Most often a fileset will be on a single tape.**
  - pnfs hierarchy of directories for file metadata.
    - After ~10,000 entries in directory access becomes innefficient.
    - 1K tapes →tapes → filesets → files stays efficient. One possible example.
- See Robert Kennedy's review for more discussion.

1st Test of Software Mapping Complete

- We've shown that DH can read / write with Enstore.
  - No change was needed to higher software: DHMods, DIM, DFC.
  - Offline policy was that our software should be compatible with Enstore.

- Read
  - Paul Hubbard modified Stager and read several filesets with Enstore.
    - Done from b0sgi02 test stand registered for STKEN use.
    - Code branched on tape label to work with Enstore or mt_tools.
      - **New "eg" label told code to use Enstore and read from STKEN robot.**
      - **Other tape labels tell code to use mt_tools to read from ADIC robot.**
  - Got full 10 MB/s rate expected for 9940 drive once tape mounted.

- Write
  - Dmitri Litvintsev has written data with Enstore to public STKEN robot.
    - DFCTestTapeWrite now has command line switch to choose method.
      - **Default is mt_tools, command line switch enables Enstore.**

Serial Media Capacities & Costs

- Existing System
  - ADIC AML/2 Robot with 20,000 slots for tapes.
  - AIT-2 cartridge with 50 GB capacity (45 GB average usage)
    - $60 cost per tape.
  - AIT-2 drives read / write of 6 / 3 MB/s (partitioned).
    - $5 K each in robot. We have 32 currently.
  - Currently have connected 8 drives for raw data & farms, 11 drives for users.
    - Rough DC rate of 9 MB/s raw logging, 9 MB/s farms logging, 9 MB/s farms reading.
    - Capable of a DC rate of 66 MB/s user reading if there were no operational problems.
- Proposed System
  - 3 STK Powderhorn robots with 5000 slots each.
    - $85 K per robot on sale now (normally >$200K). We have one: Run 1 STK Silo.
  - 9940 cartridges with 60 GB capacity (60 GB average usage)
    - $78 cost per tape. Same $/GB as we are now getting with AIT-2.
  - 9940 drives read / write of 10 / 10 MB/s (unpartitioned).
    - $27 K each. See next slide for how many and when.

STK Silo Logistics

- Space in FCC is an issue for new STK Silos.
  - Home is FCC Mezzanine currently under construction.
    - Completion scheduled between Dec. 5 & Dec. 14 with some risk.
    - D0 not expecting to get in their Silo, now at lab, until ~February.
- There are currently 2 other existing STK Silo's
  - Public STKEN silo with 2700 free slots on 2nd floor of FCC.
    - We have an account and 10 tapes and are practicing read / write.
  - CDF Run 1 silo with ~ 1-2 TB on 200 & 800 MB tapes.
    - Right next to the STKEN silo.
    - STKEN servers could then service CDF silo.
    - Need to get out run 1 data first.

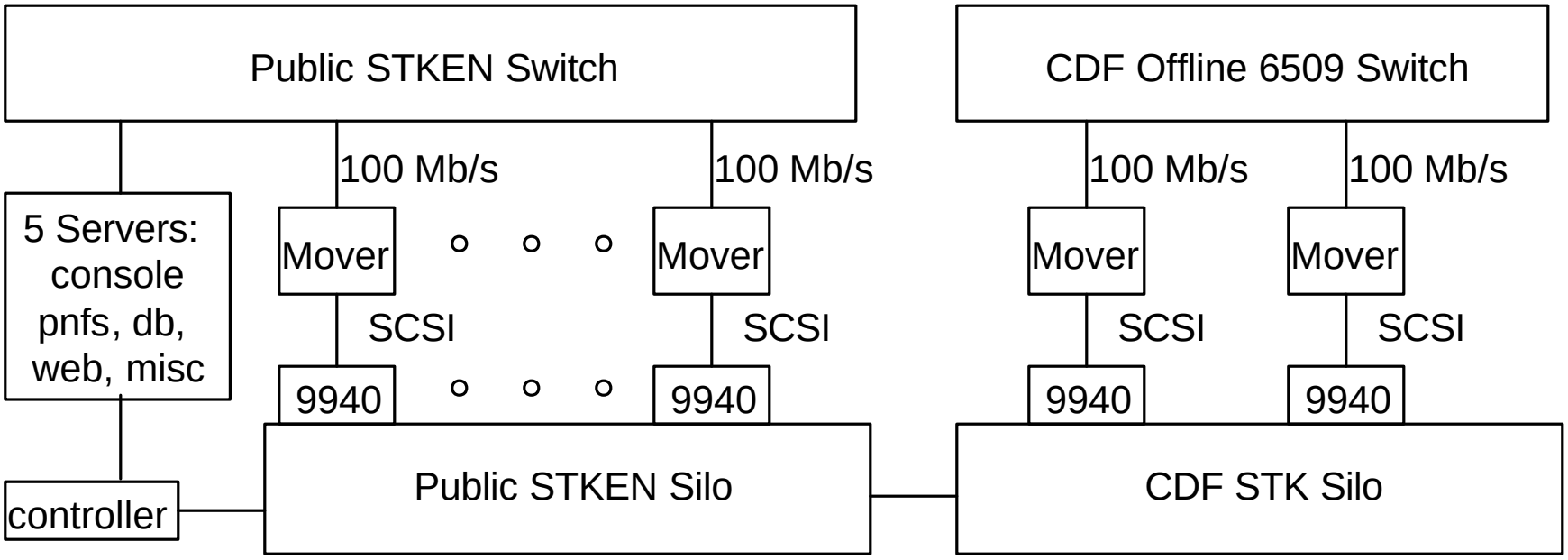
Staged Proposal

Stage	Arrival Date	Purpose	Drives & Rate	Robot	Cost & Approval
Prototype	Mid-Dec 2001 to Jan 1, 2002	Testing	2 Drives 20 MB/s	Run 1 STK Silo. 300 TB capability Share STKEN servers for admin.	$65 K Requisition signed on my approval.
Stage 1	As early as Late Jan 2002 If I start now.	DC Data rate: 30 MB/s raw & farms. 70 MB/s users	10 Drives 100 MB/s	Run 1 STK Silo. 300 TB capability CDF servers for admin.	$220 K more. I approve. Waiting for your consent.
Stage 2	As early as May 2002 ? If approved	Full Run 2A: 60 MB/s raw & farms. 150 MB/s user	21 Drives 210 MB/s	Run 1 STK Silo + 2 more STK Silos. 900 TB capability	Another $470 K more. Waiting for approval.

FY2002 Budget: $2.4 million equipment, $500K tapes.

Quick Prototype for Tape Handling

- Re-use the existing CDF Run 1 Silo and STKEN servers.
  - $65K order for 2 drives and silo renovation could be ready ~ Dec 9.
    - Potential 2-3 week slippage due to requisition & delivery delays.
  - Two mover nodes from cdf farms, takes ISD 3 days to setup.
    - Prototype hardware should be ready between Dec.12 and Jan. 1.
  - CDF should have a prototype capable of 20 MB/s & 300 TB on Jan. 1.
    - Higher rates from adding more drives & network capability (see full proposal).



Run 1 Silo Migration Plans

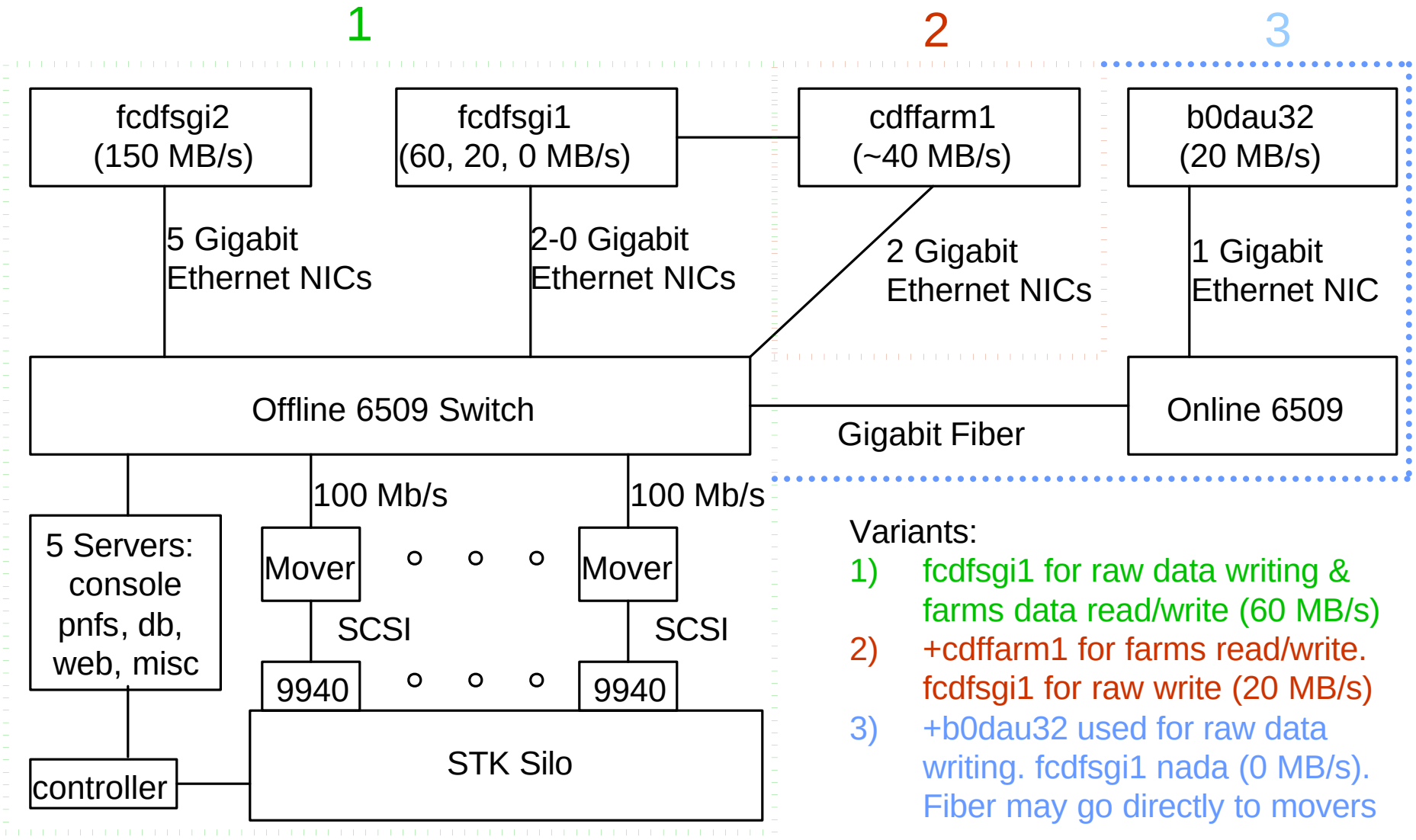
- Run 1 data will be copied to disk on cdfsga.
  - Expect to add 1-2 TB before thanksgiving.
  - Added to Run 1 Silo's disk staging pool.
    - Currently 130 GB.
  - Copy entire sample of silo data onto disk.
    - User's will access silo data as before, only it will be on disk not tape.
    - Data should be able to be copied in less than 3 weeks.
      - **Greater than ~ 1 MB/s rate out of silo for 1 – 2 TB of data.**

- Run 1 data can be copied to public STKEN robot.
  - Permanent copy that can be accessed via network.
  - After CDF Silo upgraded we can moves tapes back.

Network Issues

- Offline LAN has plenty of capacity.
  - CISCO 6509 switches can handle 32 Gb/s.
- Bandwidth from DH nodes to 6509 needs increase.
- 30 MB/s per gigabit ethernet Network Interface Card.
  - Gigabit NIC connects DH nodes to switch.
  - Requires one CPU per NIC dedicated to network traffic.
  - Routing tested on SGIs by D0 and ISD department.

Network Design of Variants



Variants:

1) fcdfsgi1 for raw data writing & farms data read/write (60 MB/s)
2) +cdffarm1 for farms read/write. fcdfsgi1 for raw write (20 MB/s)
3) +b0dau32 used for raw data writing. fcdfsgi1 nada (0 MB/s). Fiber may go directly to movers

Discussion of Network Variants

- fcdfsgi1 used for raw data logging & farms staging.
  - Same as now, so minimal change to operations (+).
  - May require 1-2 extra CPUs on fcdfsgi1 (-).
    - 4 available now, and we would dedicate 2 to Gigabit connections.
  - fcdfsgi1 single point of failure for raw data & farms (-).
- cdffarm1 SGI used for farms staging (reading & writing).
  - Farms input / output under control of farms group (+).
  - May require additional CPUs on cdffarm1 as above (-)
  - Farms can process data independent of fcdfsgi1 (+).
- Online b0dau32 SGI used for raw data logging
  - Cleanest way to log raw data with Enstore. What D0 does. (+)
  - Removes need for CXFS filesystem that causes operational problems (+).
  - Leaves fcdfsgi1 untouched for use of legacy system. Clean migration. (+)
  - Raw data logging under control of online group, likely Rochester (+).
  - May require 1 additional CPU for b0dau32 as above (-).
  - Plan crosses organizational boundaries at lab and collaboration (-).

Installation Schedule

- Prototype between Mid-Dec and Jan.1
  - 4-6 weeks: Run 1 silo upgraded to support two 9940 drives.
  - 3 days: attach two mover nodes.
  - We began this process on Nov. 9.
- Stage 1 hardware in as little as 8 weeks.
  - 6 weeks: requisition, delivery & installation for 8 additional drives.
  - 2 weeks: ISD installs mover nodes.
  - In parallel with these 8 weeks
    - ISD installs server nodes dedicated to CDF.
    - Obtain CPUs & gigabit NICS for fcdffsgi1 / 2, cdffarm1, b0dau32.
    - Attach network infrastructure.
  - If approved now a system could be installed in late Jan. for $220K
- Stage 2 hardware in as little as 12 weeks
  - Two more robots, 11 more drives, install mover nodes & server nodes.

Use & Migration Options

- DH software will support both Enstore / mt_tools

- Option 1: Raw data logging last
  1. Begin copying data in ADIC robot to prototype in Jan.
     Takes 1 ½ months ito copy all data using 5 AIT-2 drives.
  2. Users read / write secondary datasets to prototype in Jan.
     Takes a few weeks to shakedown user problems.
  3. Farms reads copied data from Stage 1 robot in late Jan.
     Takes a few days because they were testing on prototype in Jan.
  4. Farms writes to Stage 1 robot in early Feb.
     Full test of farms reading and writing completed.
  5. Raw data logged to Stage 1 robot in late February.
     All activities switched to using Stage 1 robot by March.

Use & Migration Options

- Option 2: Raw Data Logging ASAP.
  1. Copy raw data to prototype beginning in Jan.
     Takes about 1 ½ months to complete for 5 AIT-2 drives.
  2. Test users read and write data to prototype in Jan.
     Takes a few weeks to shakedown problems.
  3. Log raw data to stage 1 robot in parallel with ADIC robot.
     If possible to orchestrate, could begin in late Jan.
  4. Farms read/write to stage 1 robot in parallel with ADIC.
     If possible to orchestrate, could begin in late Jan.
  5. Turn off writing to ADIC robot.
     Occurs roughly one month after system available. Late Feb.
  6. Turn off reading from ADIC robot.
     As soon as copy is complete, as early as March.

Conclusions

- Existing system has significant operational problems.
  - We cannot guarantee it will run under a larger load in the future.
- Robust CD supported systems are in use at FNAL.
  - Enstore network access to STK 9940 drives in STK robots.
  - Used by 16 experiments & supported by ISD department.
  - Fits in well with plans for CAF, GRID, and analysis in the trailers.
- We have started on a plan for a prototype.
  - $65 K will provide some capability as early as middle of Dec.
- We have a plan for a full system
  - Stage 1: $220 K more from our $2.4 M budget for DC operations by March.
  - Stage 2: $170 K more for 2 STK silos and $300 K for 11 more drives ultimately.
- The committee should recommend on Stage 1 ASAP.
  - Prefer collaboration to agree on this direction before I spend a lot more money.